Deploying Analytics Pipelines

Session 1

Brock Tibert

BU

Boston University Questrom School of Business

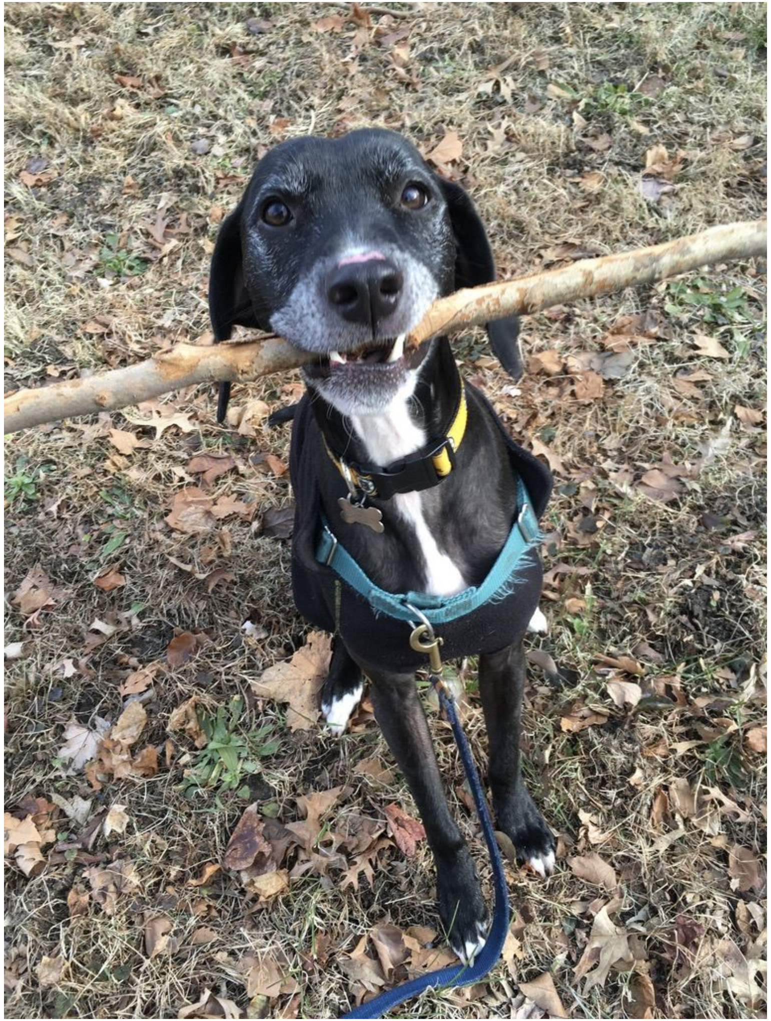
Welcome! Overview for Today

- About me
- Why this course
- Class introductions
- Course delivery and expectations
- Tooling Review
- Next up: Data Modeling/ERD/DDL and DML

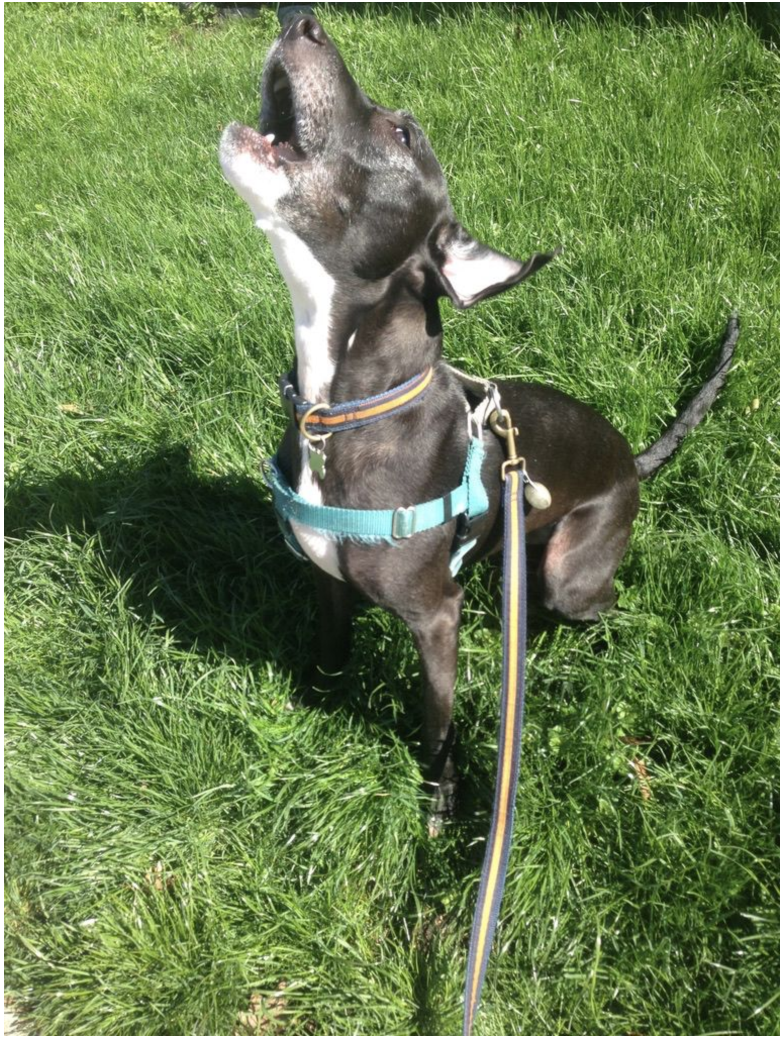
About Me

- 7th academic year at Questrom in the IS Department
- Coming on 20 years working with (and deploying) data tasks (analytics/ml, pipelines, products)
- Advise product and data teams
- Interests are in applied data, building data products.
- My areas of interest are education and sports analytics.

My Dog Bodhi (Most of the time)

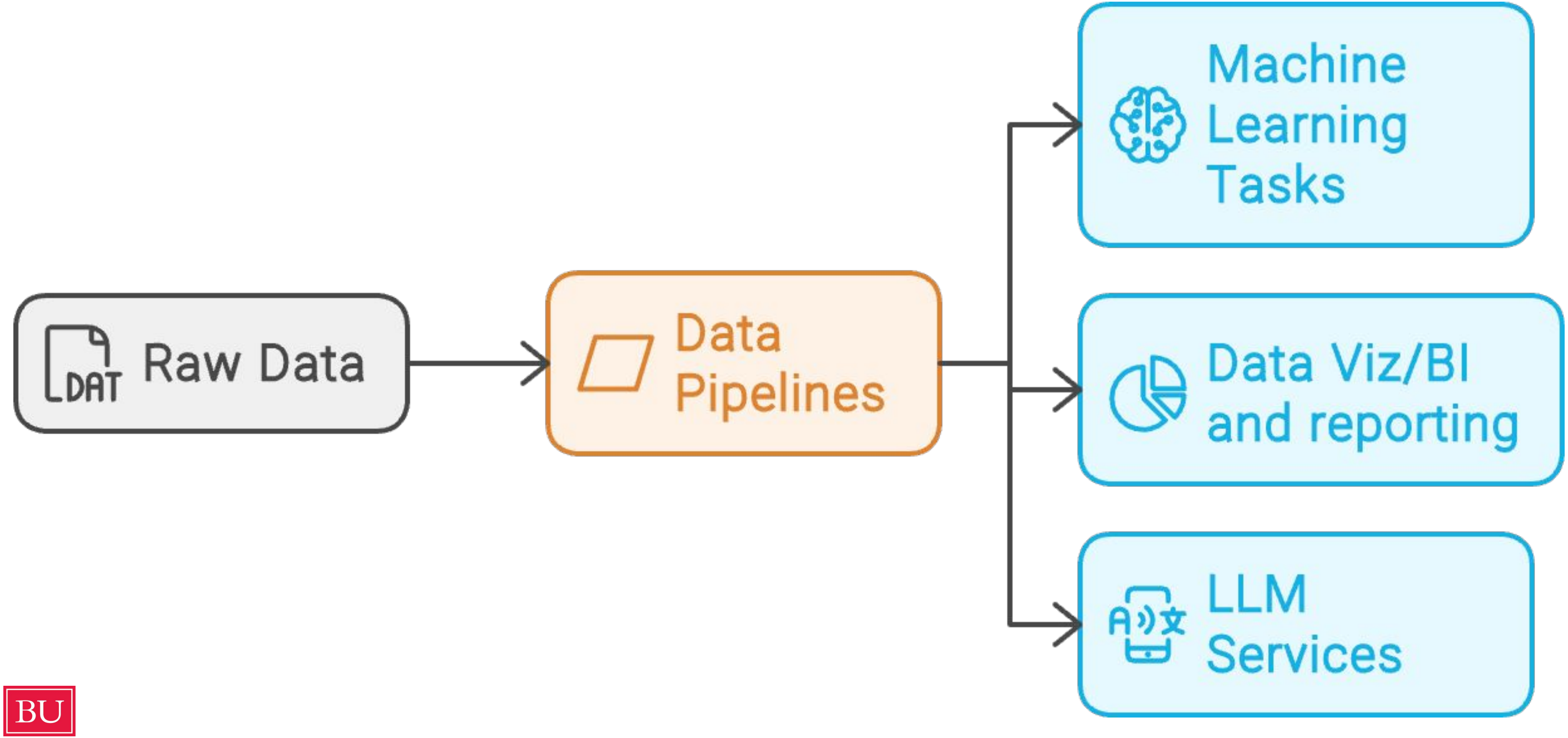


My Dog Bodhi (*when I record Supplemental Videos for you*)



Why I designed this course

Framing my experiences into three branches



Raw Data

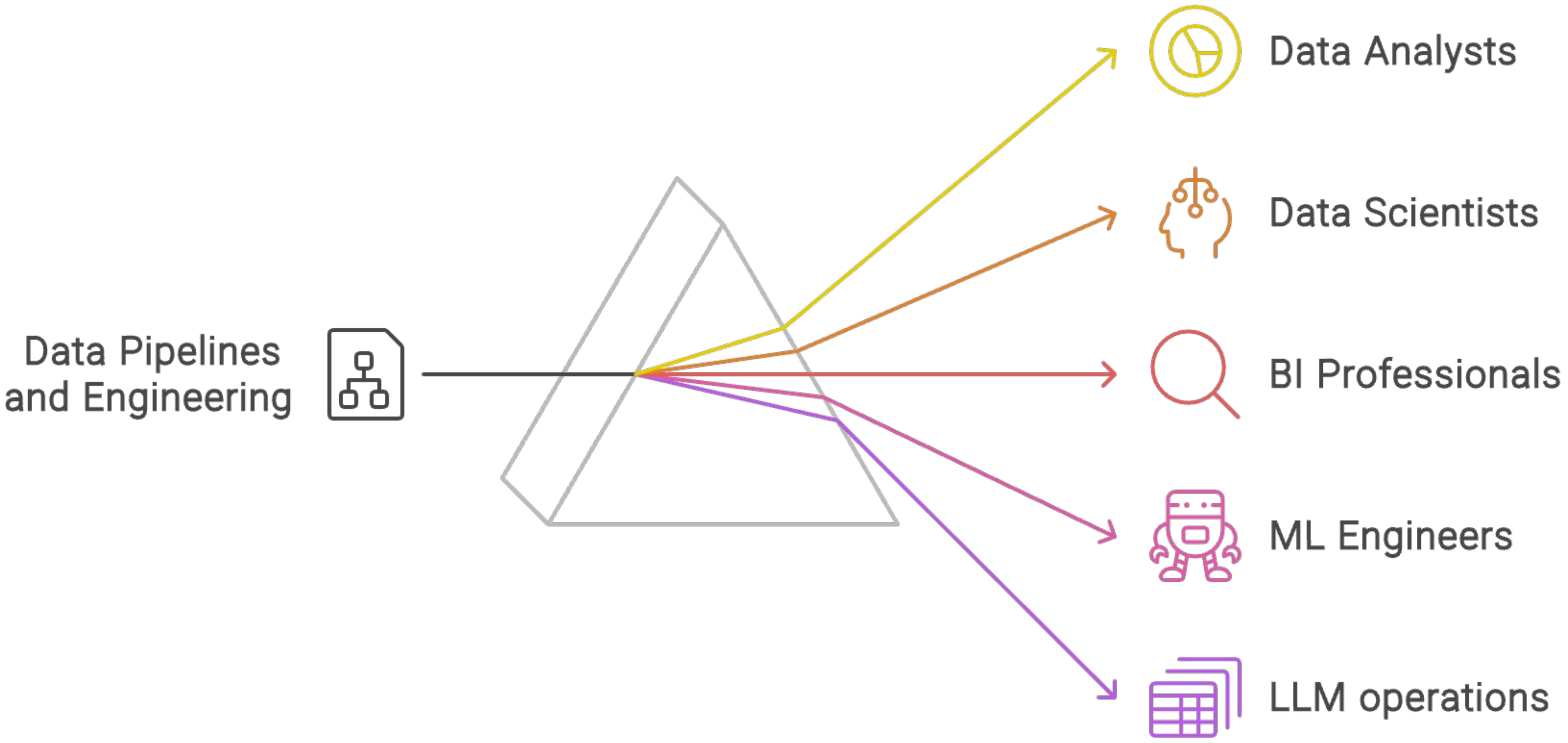
Data Pipelines

Machine Learning Tasks

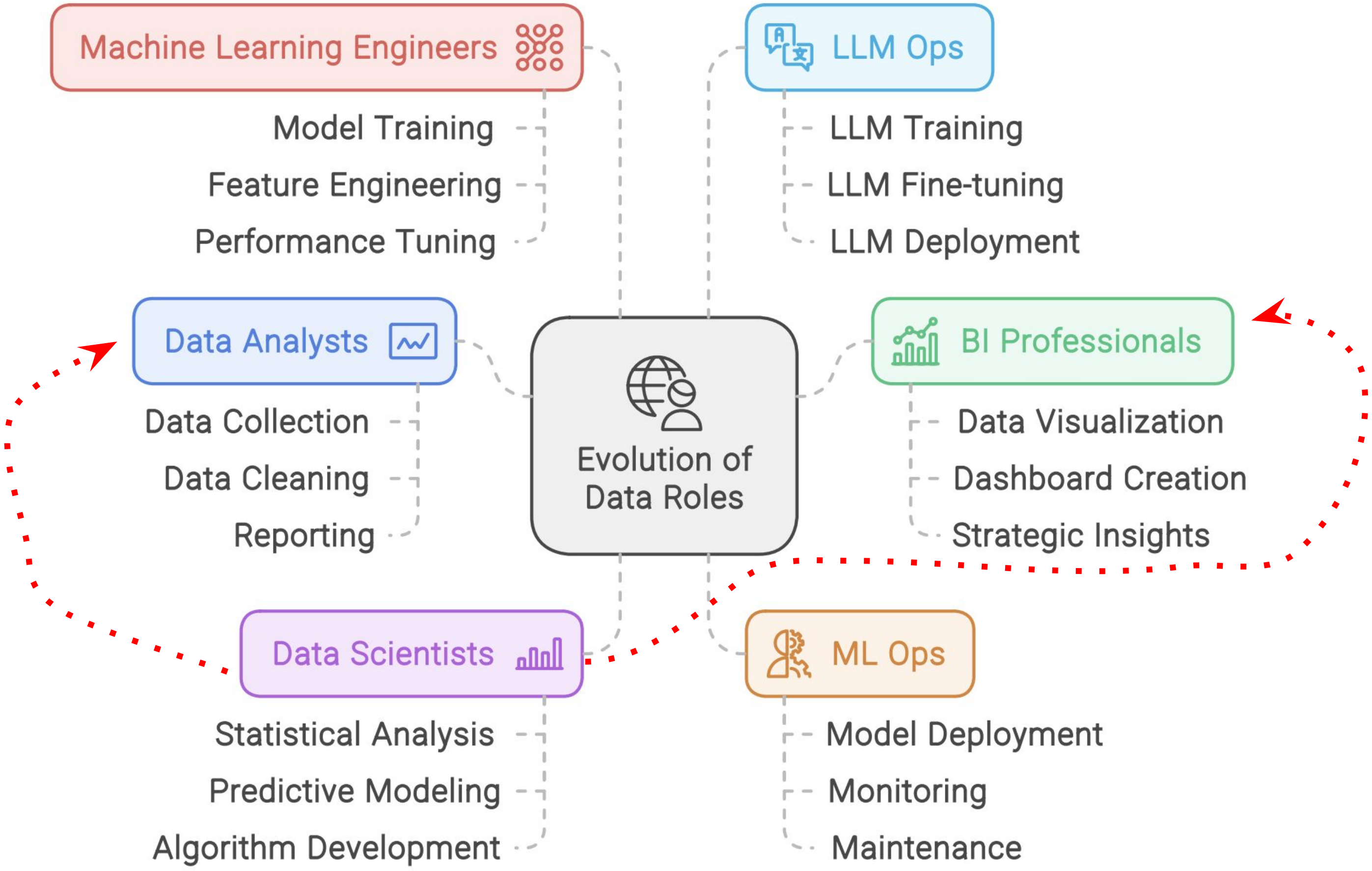
Data Viz/BI and reporting

LLM Services

My view: “data teams”


Data Pipelines and Engineering
Data Analysts
Data Scientists
BI Professionals
ML Engineers
LLM operations

My View: Evolution of data roles and data teams



Evolution of Data Roles

Machine Learning Engineers

- Model Training
- Feature Engineering
- Performance Tuning

LLM Ops

- LLM Training
- LLM Fine-tuning
- LLM Deployment

Data Analysts

- Data Collection
- Data Cleaning
- Reporting

BI Professionals

- Data Visualization
- Dashboard Creation
- Strategic Insights

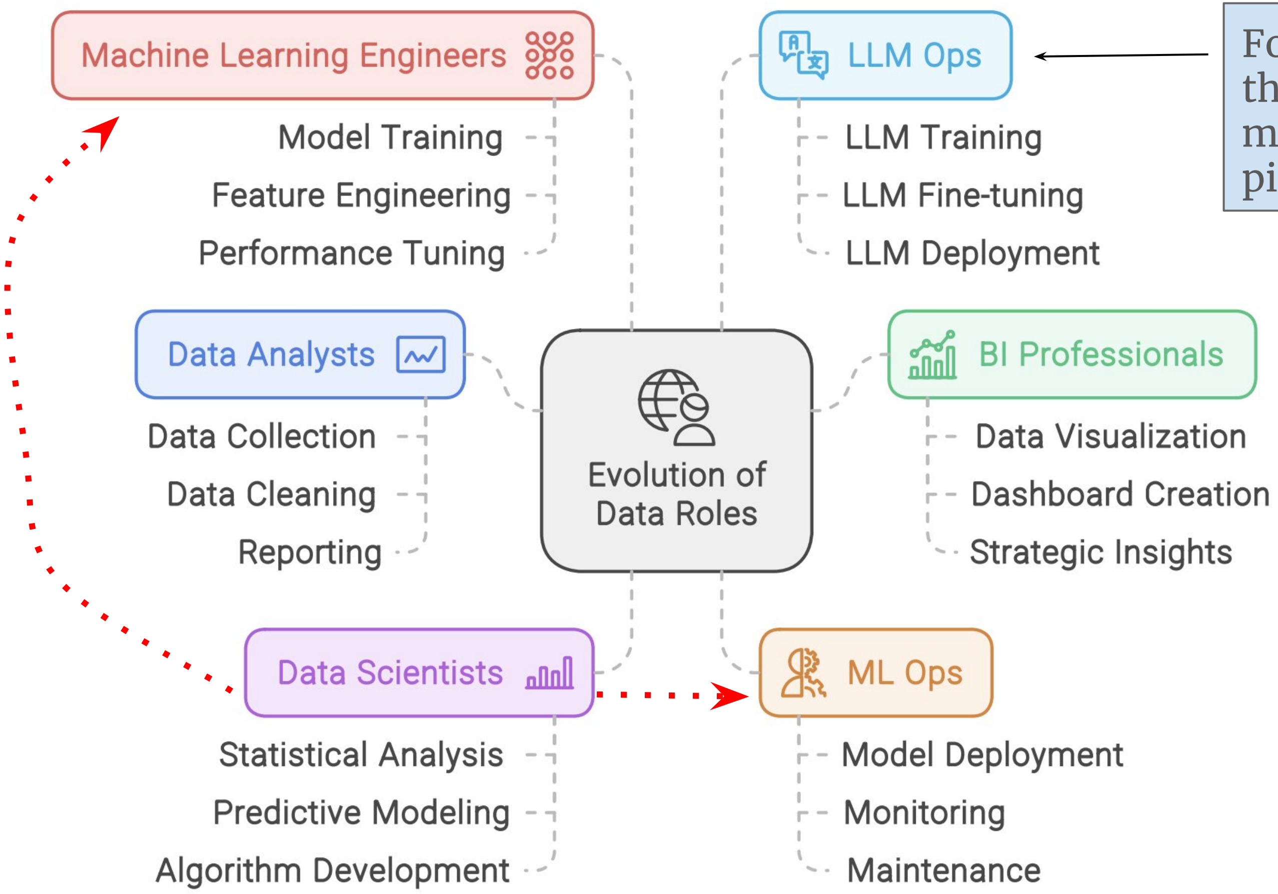
Data Scientists

- Statistical Analysis
- Predictive Modeling
- Algorithm Development

ML Ops

- Model Deployment
- Monitoring
- Maintenance

My View: Evolution of data roles and data teams



Evolution of Data Roles

Machine Learning Engineers
- Model Training
- Feature Engineering
- Performance Tuning

LLM Ops
- LLM Training
- LLM Fine-tuning
- LLM Deployment

Data Analysts
- Data Collection
- Data Cleaning
- Reporting

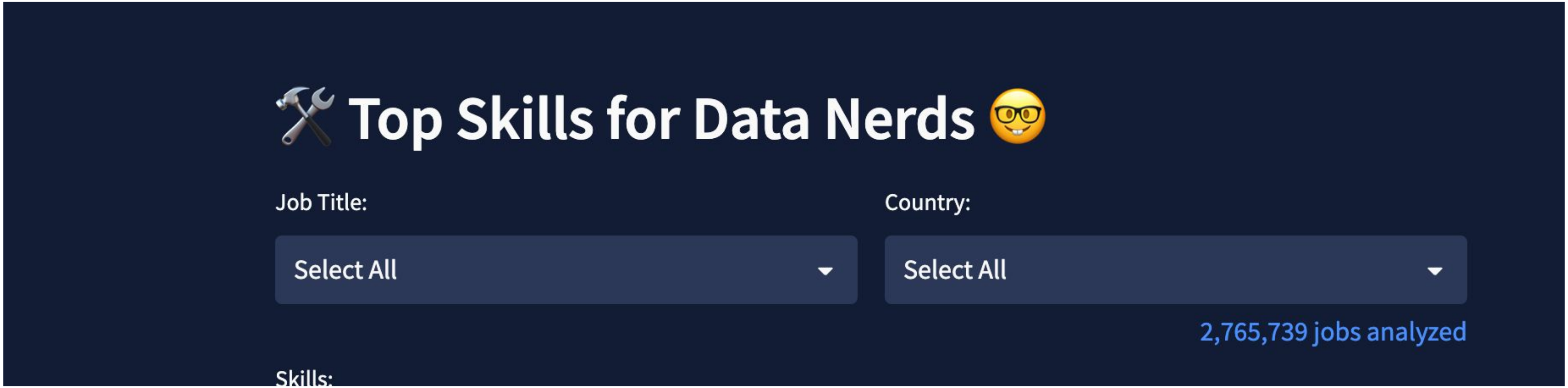
BI Professionals
- Data Visualization
- Dashboard Creation
- Strategic Insights

Data Scientists
- Statistical Analysis
- Predictive Modeling
- Algorithm Development

ML Ops
- Model Deployment
- Monitoring
- Maintenance

For most shops, this likely will be more focused on pipelines

A Different View:
Class Resources > Analytics Persona



🛠️ Top Skills for Data Nerds 🤓

Job Title:

Country:

2,765,739 jobs analyzed

Skills:

Class Intros

1. About you
2. What are you looking to get out of this course?
3. What are you looking to do after Questrom?

Course Delivery Discussion

Cloud Resources

Google Cloud

Course Delivery

- Three major topics that are integrated and build up as we go.
- The delivery of the material will *roughly* follow this pattern
  - First half is a discussion of the core themes for the day, and demos
  - Second half will be your turn for hands on practice to review the code/patterns applied
- You may want to take notes and watch when I am "live coding" or perform demonstrations in class
- I wrote this course based on the what I know you covered in the prerequisite courses. My responsibility is to help you extend the foundational knowledge into how we can move from our laptop to production
- I will try to be flexible in the delivery of this course. **I care about your learning above all else**, so I will supplement and tweak the arc/material as needed.

Course Expectations –> Syllabus Review

Course Expectations – Cont’d

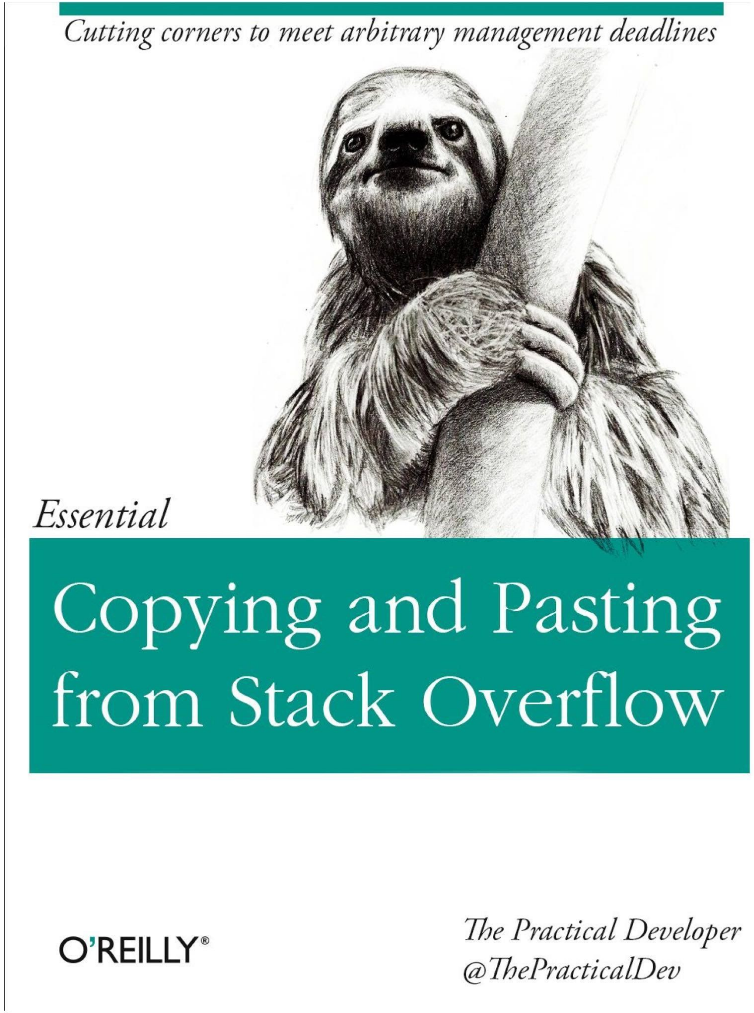
- Feel free to ask questions along the way (this is intended to be an interactive class). You are expected to be an active participant.
- This is a **safe space**. Data are hard, and we will all help each other.
- You may not know all of the answers, that is ok. **You should be here to sharpen your data skills by building experience beyond your laptop**
- You are expected to come to class prepared.
  - Making progress on your projects
  - Building intuition of the core concepts and thinking critically about how to apply them
  - Practice, tutorials, trial-and-error hobby projects, DataCamps.
  - **Only You** know best how you need to prepare each week.
- You will get a break (approx.) halfway through each class.
- You are expected to treat everyone with respect
- Being able to work through challenges is part of this course. I won’t give you all of the answers, but I will help you work through the problems.

Teamwork and Assessments

- I will assign the teams before next week.
- Team-based deliverables are a core component of learning in this class
  - The three deliverables are intentional to help you simulate how a data teams tasks incrementally evolve
- You will be responsible for selecting the data **feeds** and tasks that you will deploy this semester. I am here to mentor you and guide you as you work through those deliverables
- I feel very strongly that working in teams during this course will help your onboarding of a data team after Questrom
- I will use a variety of assessments to help you with the core learning themes, and will also challenge you to think through arbitrary problems that you will face after Questrom as it relates to deploying analytics pipelines

Blackboard → Team Project Ideas
Let’s review

I used to show this slide in my coding classes.



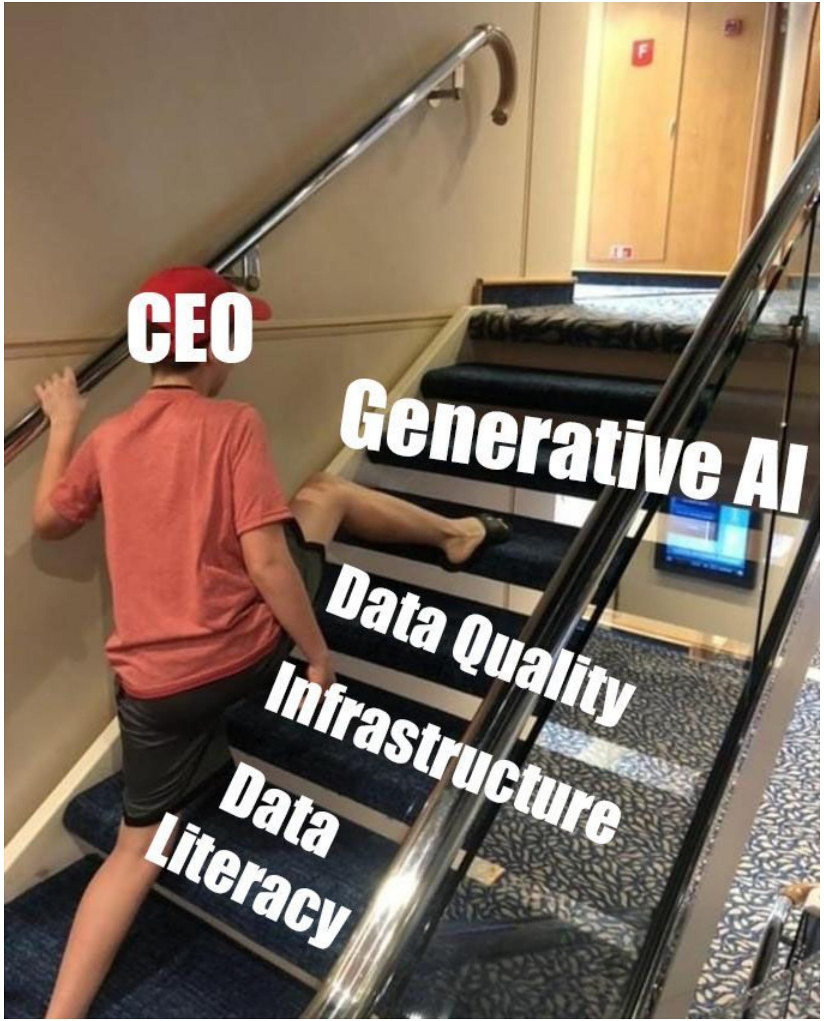
Cutting corners to meet arbitrary management deadlines

Essential

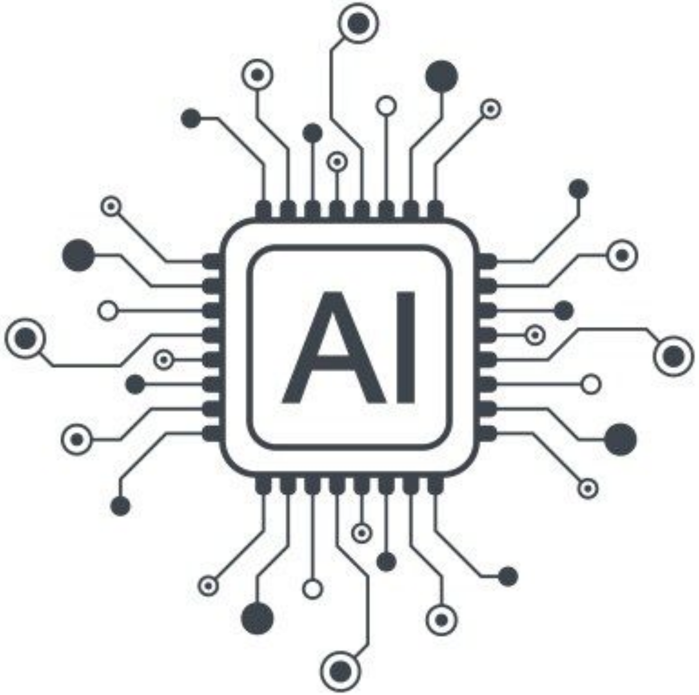
Copying and Pasting from Stack Overflow

O'REILLY®

The Practical Developer
@ThePracticalDev


CEO
Generative AI
Data Quality
Infrastructure
Data
Literacy

execs in 2023
literally any app
ChatGPT


AI

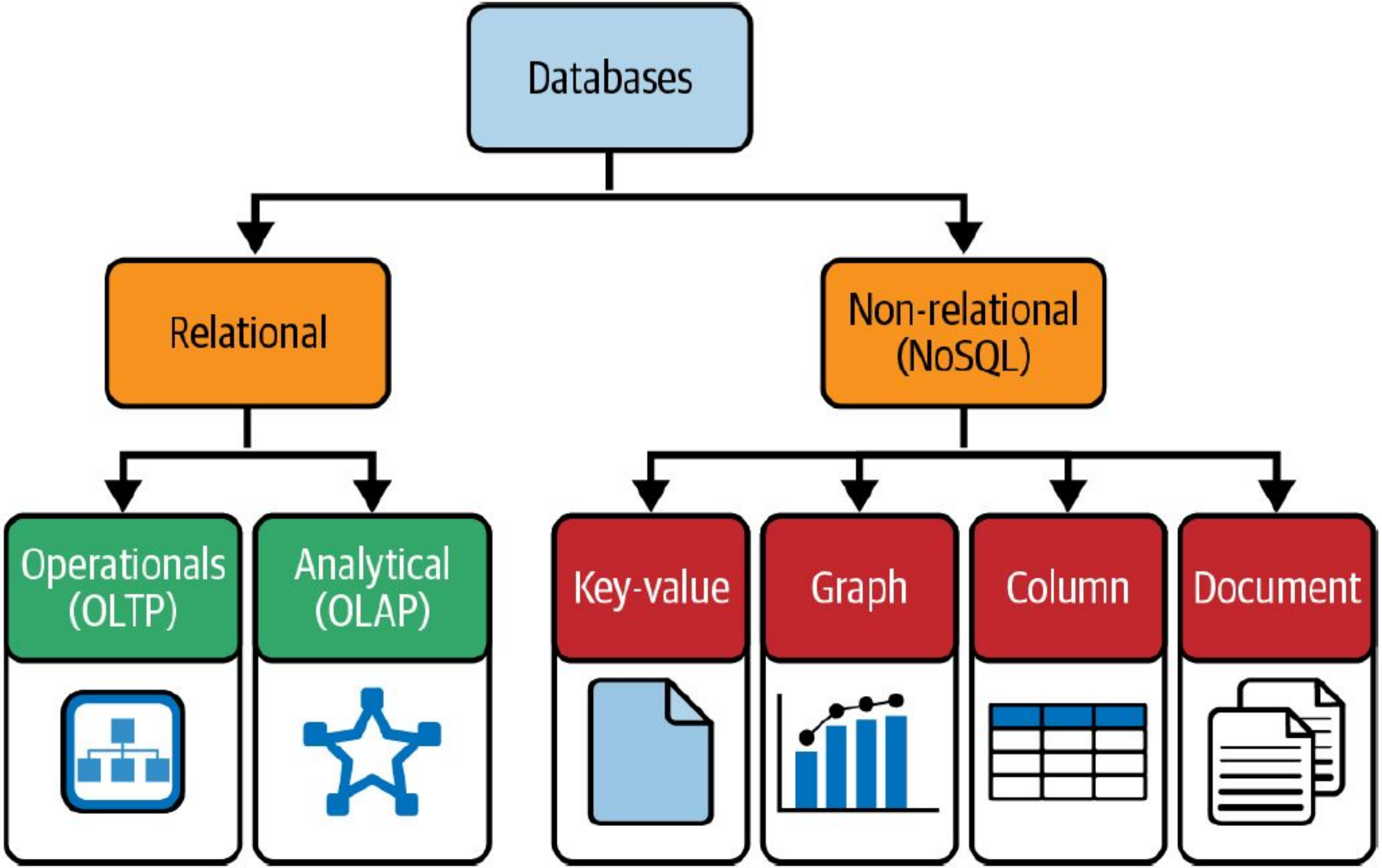
Some logistics

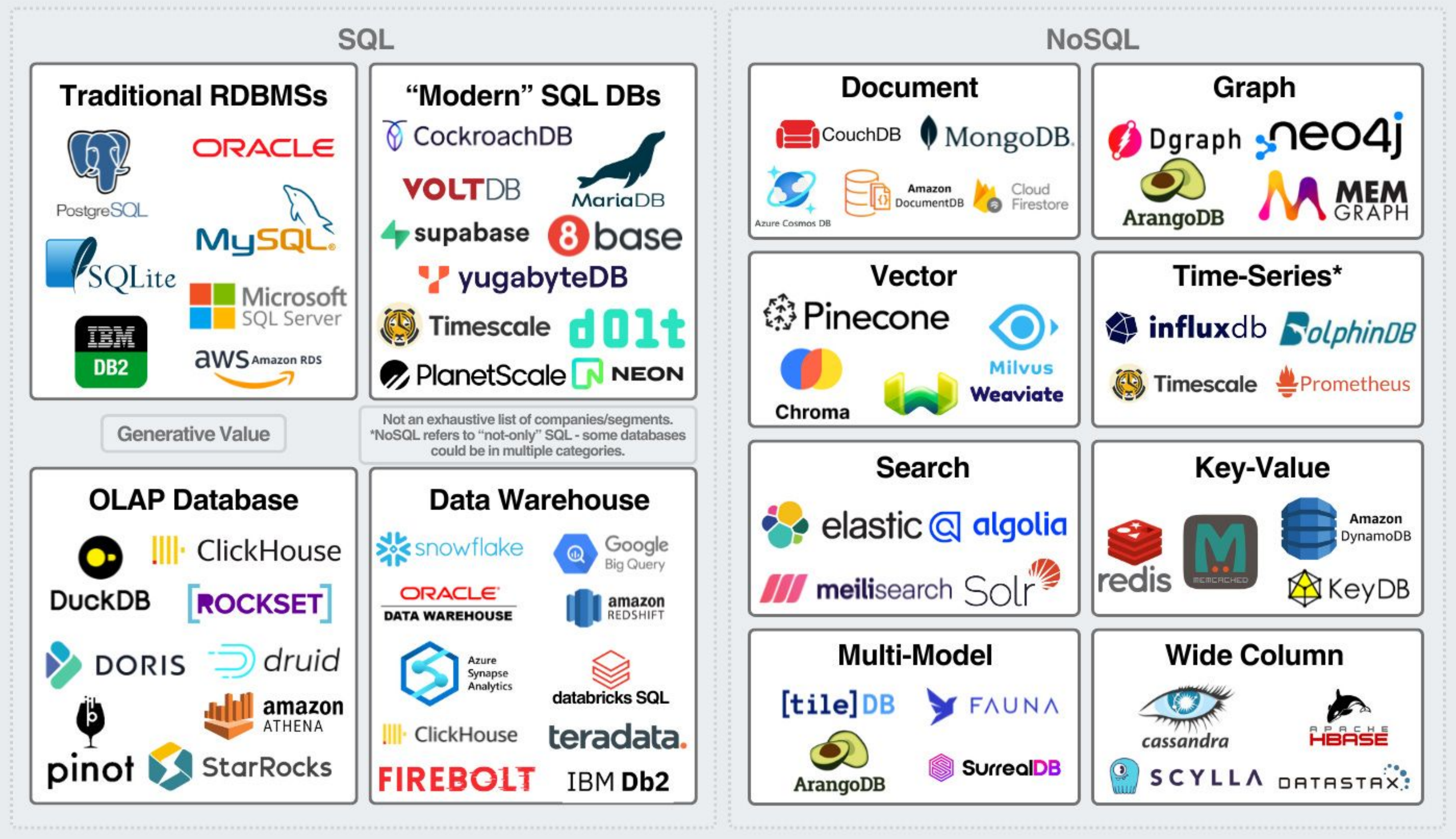
datacamp

Attendance: Arkaive

- Instructions on Blackboard
- Can download app or use web browser
- Check-ins will be periodic and the first 15 minutes of class.
- We will start on time, and have a break halfway through each class session.

Analytics start with the database, right?


Databases
Relational
Non-relational (NoSQL)
Operationals (OLTP)
Analytical (OLAP)
Key-value
Graph
Column
Document



SQL

Traditional RDBMSs

- PostgreSQL
- ORACLE
- MySQL
- SQLite
- Microsoft SQL Server
- IBM DB2
- aws Amazon RDS

"Modern" SQL DBs

- CockroachDB
- VOLTDB
- MariaDB
- supabase
- 8base
- yugabyteDB
- Timescale
- d0lt
- PlanetScale
- NEON

Generative Value

Not an exhaustive list of companies/segments.
*NoSQL refers to "not-only" SQL - some databases could be in multiple categories.

OLAP Database

- DuckDB
- ClickHouse
- ROCKSET
- DORIS
- druid
- pinot
- amazon ATHENA
- StarRocks

Data Warehouse

- snowflake
- Google Big Query
- ORACLE DATA WAREHOUSE
- amazon REDSHIFT
- Azure Synapse Analytics
- databricks SQL
- ClickHouse
- teradata.
- FIREBOLT
- IBM Db2

NoSQL

Document

- CouchDB
- MongoDB.
- Azure Cosmos DB
- Amazon DocumentDB
- Cloud Firestore

Graph

- Dgraph
- neo4j
- ArangoDB
- MEM GRAPH

Vector

- Pinecone
- Milvus
- Chroma
- Weaviate

Time-Series*

- influxdb
- DolphinDB
- Timescale
- Prometheus

Search

- elastic
- algolia
- meilisearch
- Solr

Key-Value

- redis
- MEMCACHED
- Amazon DynamoDB
- KeyDB

Multi-Model

- [tile]DB
- FAUNA
- ArangoDB
- SurrealDB

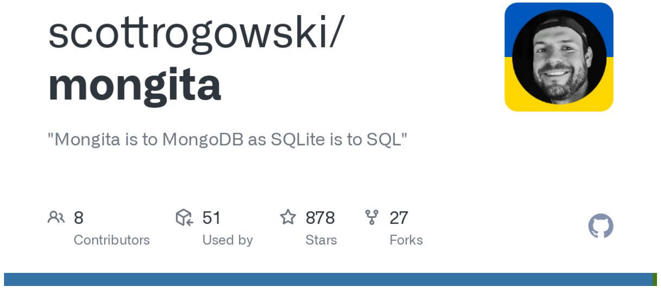
Wide Column

- cassandra
- APACHE HBASE
- SCYLLA
- DATASTAX

But wait, there’s more: In-Process/Embeddable Database Options

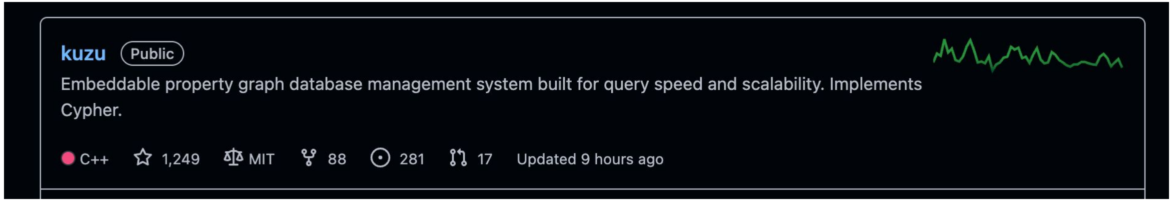
SQLite

DuckDB


scottrogowski/
mongita

"Mongita is to MongoDB as SQLite is to SQL"

8	51	878	27
Contributors	Used by	Stars	Forks

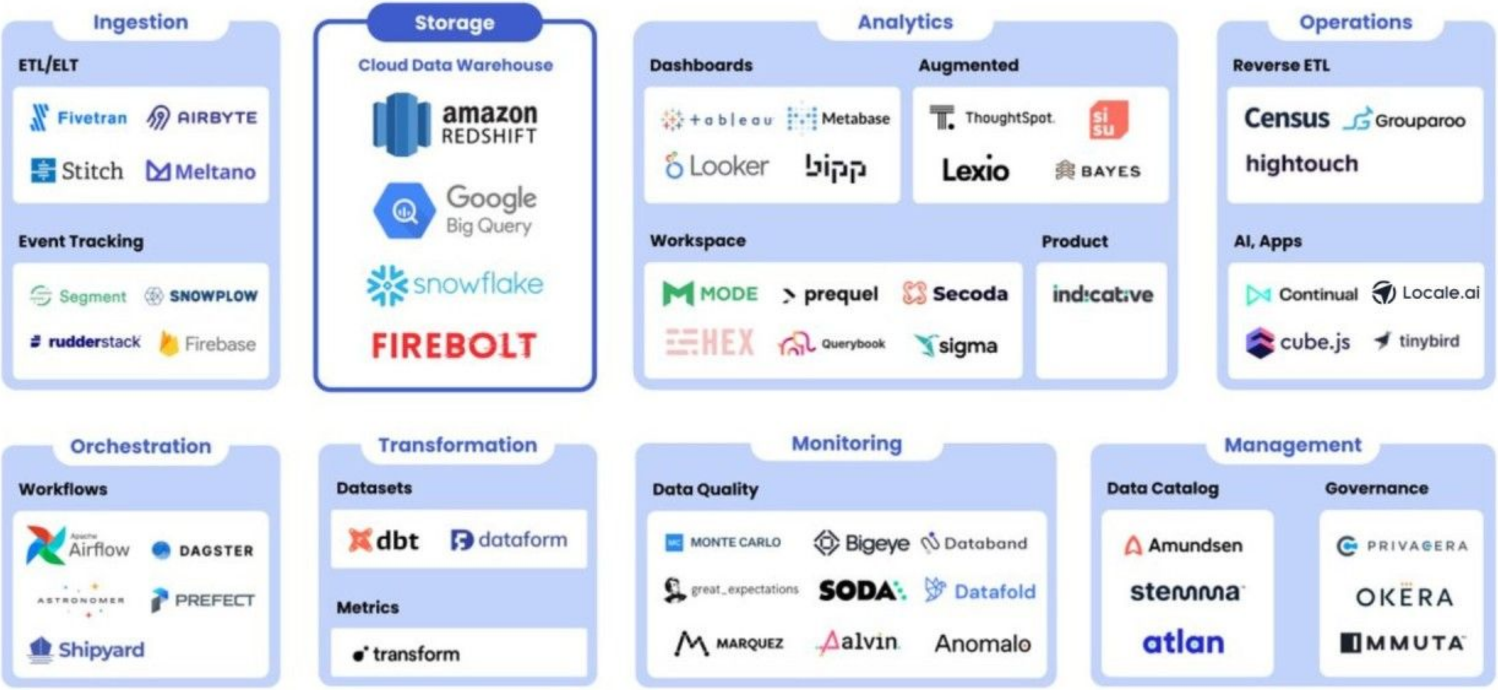

kuzu Public

Embeddable property graph database management system built for query speed and scalability. Implements Cypher.

● C++ ☆ 1,249 MIT 88 281 17 Updated 9 hours ago

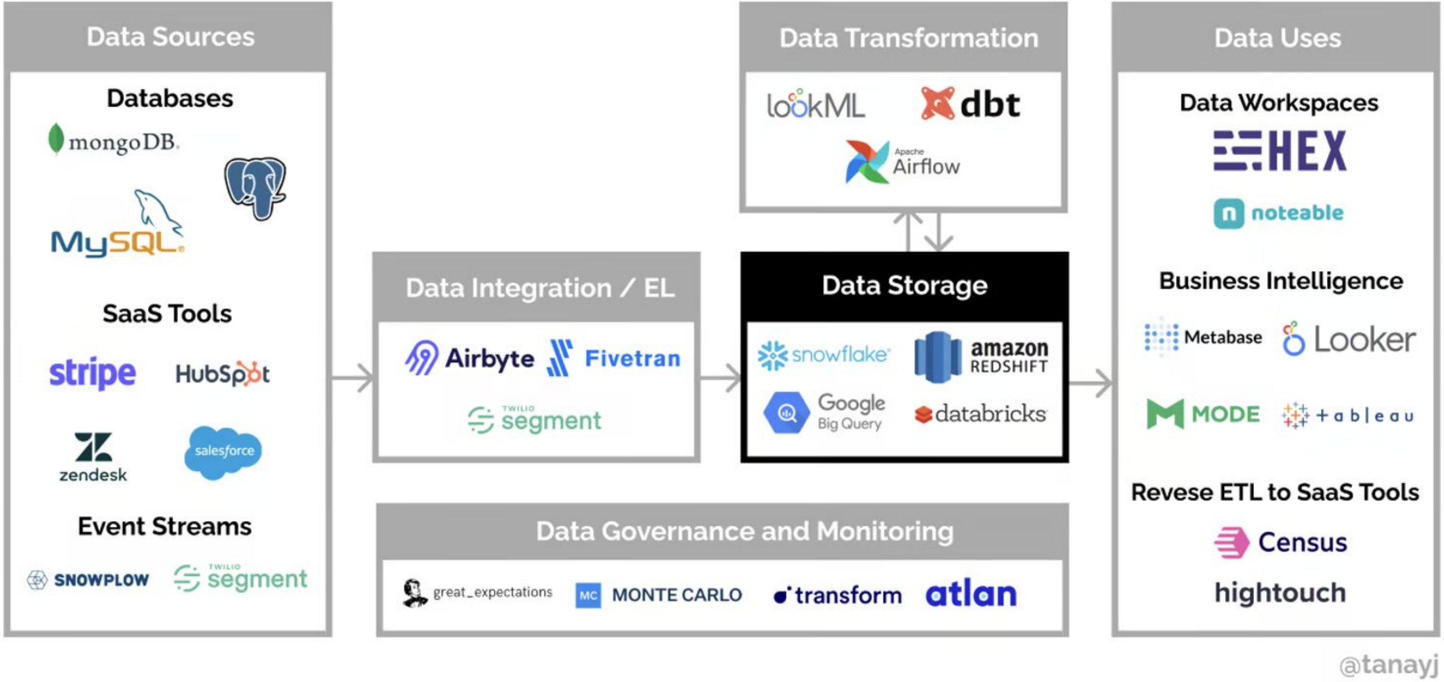
The “modern data stack”

Modern Data Stack – Example 1

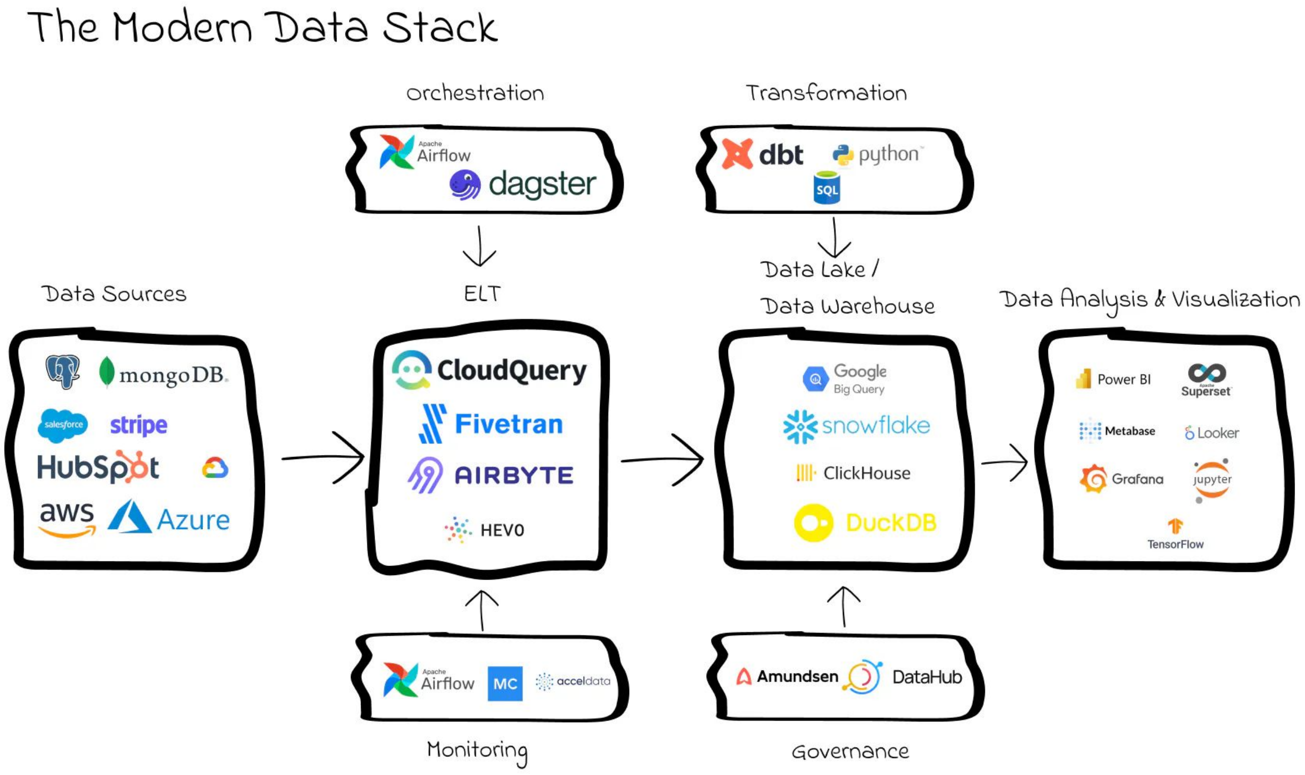


Boston University Questrom School of Business

Modern Data Stack – Example 2



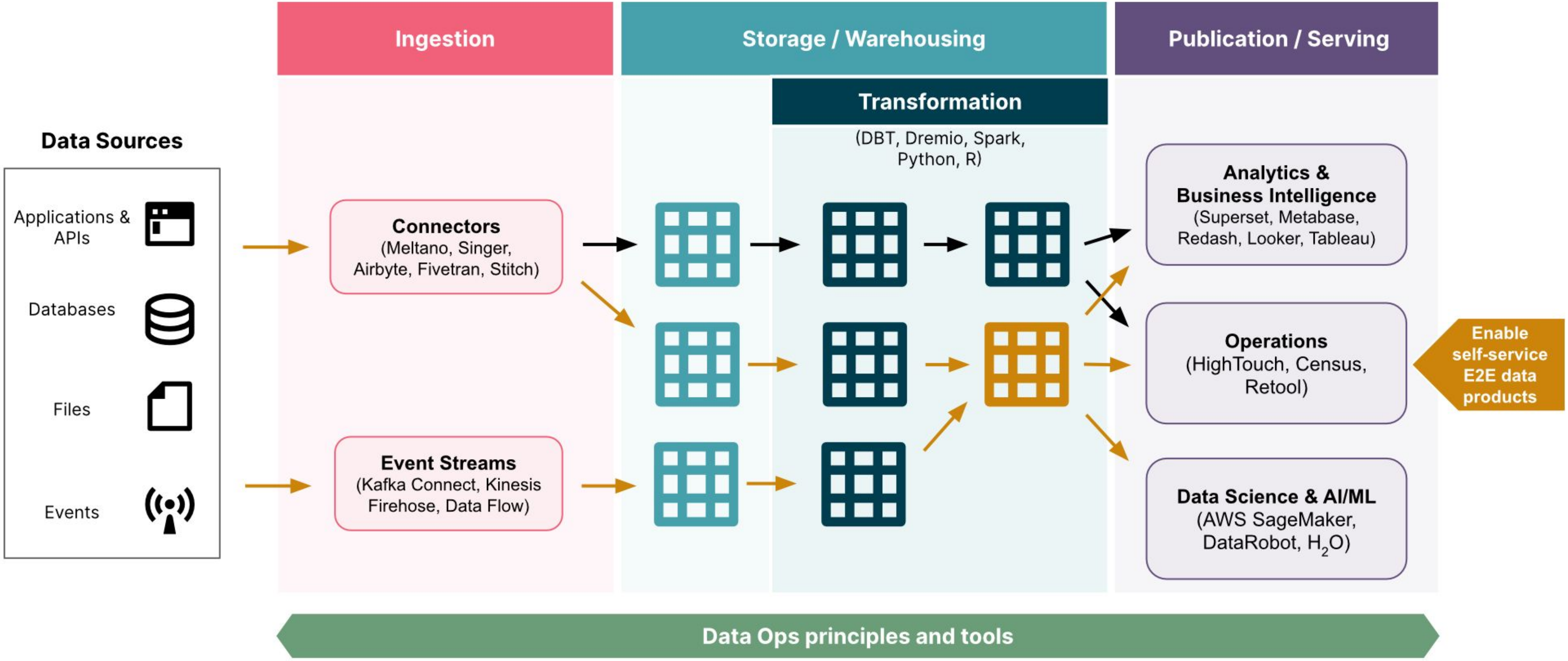
Modern Data Stack – We can start to see *some* patterns



By CloudQuery

Boston University Questrom School of Business

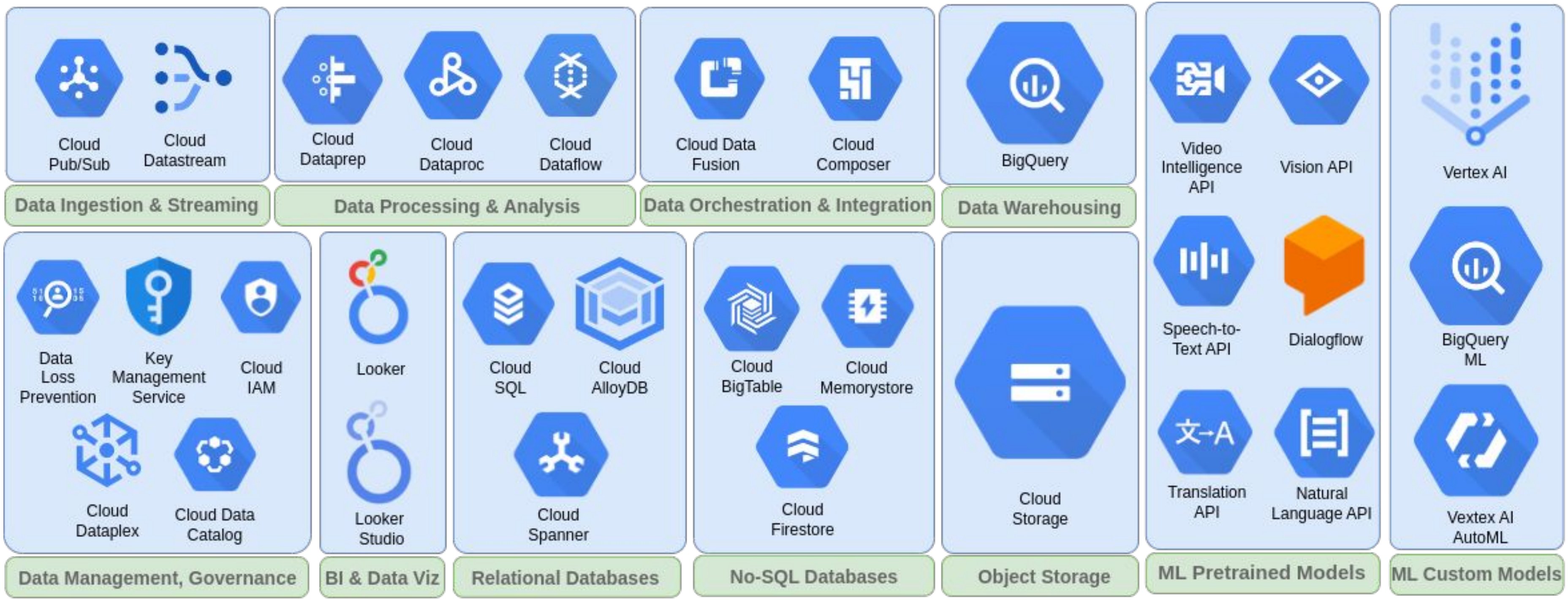
Modern Data Stack – A reduction in some of the noise



Ingestion
Storage / Warehousing
Publication / Serving
Transformation
(DBT, Dremio, Spark, Python, R)
Data Sources
Applications & APIs
Databases
Files
Events
Connectors (Meltano, Singer, Airbyte, Fivetran, Stitch)
Event Streams (Kafka Connect, Kinesis Firehose, Data Flow)
Analytics & Business Intelligence (Superset, Metabase, Redash, Looker, Tableau)
Operations (HighTouch, Census, Retool)
Data Science & AI/ML (AWS SageMaker, DataRobot, H2O)
Enable self-service E2E data products
Data Ops principles and tools

NOTE: Also starting to see some cloud hosted service references

Cloud Provider Services – GCP



Data Ingestion & Streaming
- Cloud Pub/Sub
- Cloud Datastream

Data Processing & Analysis
- Cloud Dataprep
- Cloud Dataproc
- Cloud Dataflow

Data Orchestration & Integration
- Cloud Data Fusion
- Cloud Composer

Data Warehousing
- BigQuery

ML Pretrained Models
- Video Intelligence API
- Vision API
- Speech-to-Text API
- Dialogflow
- Translation API
- Natural Language API

ML Custom Models
- Vertex AI
- BigQuery ML
- Vextex AI AutoML

Data Management, Governance
- Data Loss Prevention
- Key Management Service
- Cloud IAM
- Cloud Dataplex
- Cloud Data Catalog

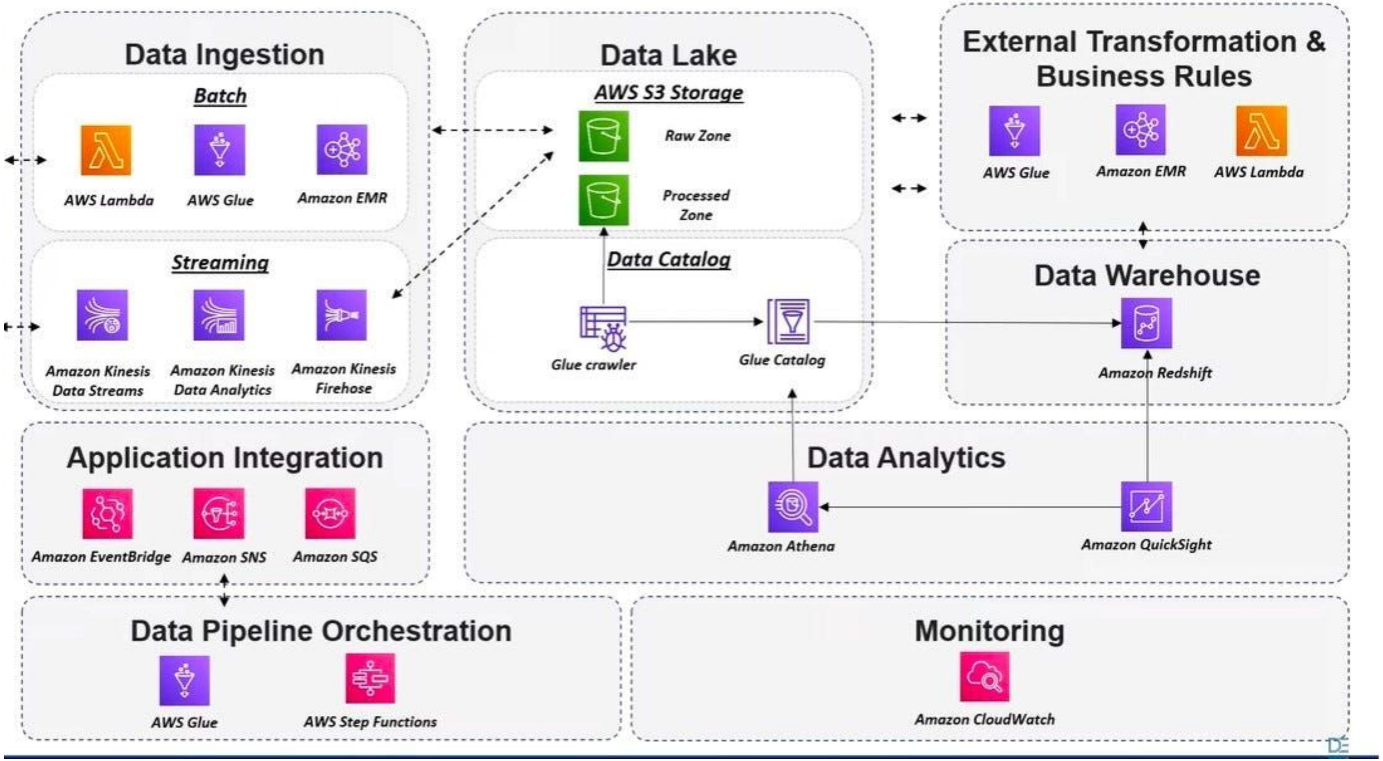
BI & Data Viz
- Looker
- Looker Studio

Relational Databases
- Cloud SQL
- Cloud AlloyDB
- Cloud Spanner

No-SQL Databases
- Cloud BigTable
- Cloud Memorystore
- Cloud Firestore

Object Storage
- Cloud Storage

Cloud Provider Services – AWS



Collaborative Thought Exercise

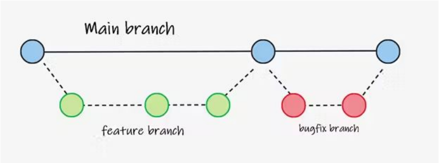
Open Source/Frameworks
vs.
Cloud-hosted Services

3 minutes: Talk with your neighbors

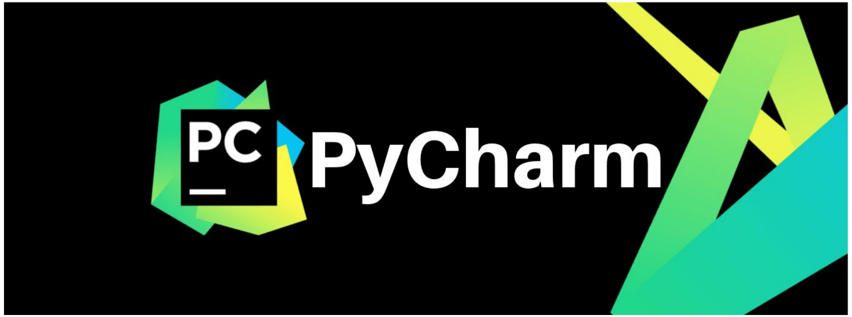
Group Notes:

- Open Source is Flexible, Data Team to keep simple via the cloud (p/c)
- Open Source docs (or lack thereof)
- Versions and new feature availability
- Scalability concerns
    - Someone leaves that was managing it
    - Team members need to know how to run
    - Does the self-managed server timeout? Run out of memory? Crash?

Tools:
No requirements here, choose the stack you prefer to work with!



Code IDE: My preference is VSC, but it's up to you



Supplemental Materials >
Document on tool setup

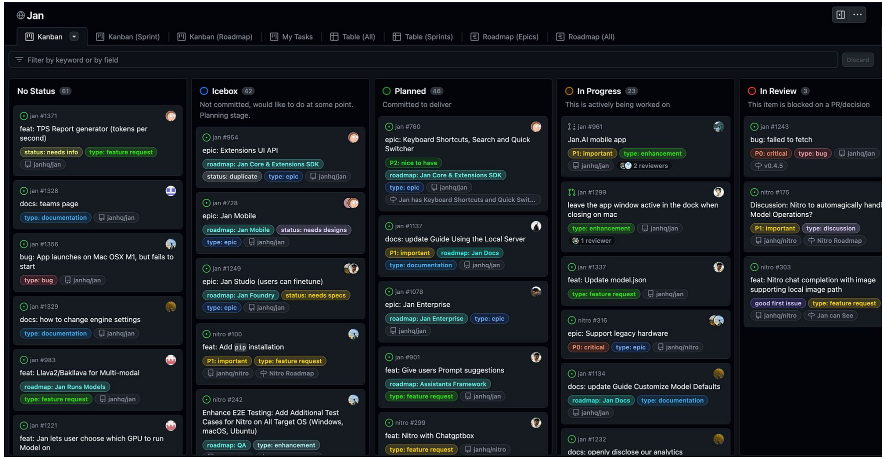
- Coding Assistant from Amazon
- Free tier eligible
- Requires separate account called builder
- Links on Blackboard

Github: Source Control and Project Management

Docker Desktop

Database IDE

Project management (Github)



Jan

Kanban | Kanban (Sprint) | Kanban (Roadmap) | My Tasks | Table (All) | Table (Sprints) | Roadmap (Epics) | Roadmap (All)

Filter by keyword or by field

Discard

No Status 61

- jan #1371
  feat: TPS Report generator (tokens per second)
  status: needs info · type: feature request · janhq/jan
- jan #1328
  docs: teams page
  type: documentation · janhq/jan
- jan #1356
  bug: App launches on Mac OSX M1, but fails to start
  type: bug · janhq/jan
- jan #1329
  docs: how to change engine settings
  type: documentation · janhq/jan
- jan #983
  feat: Llava2/Bakllava for Multi-modal
  roadmap: Jan Runs Models · type: feature request · janhq/jan
- jan #1221
  feat: Jan lets user choose which GPU to run Model on

Icebox 42

Not committed, would like to do at some point. Planning stage.

- jan #954
  epic: Extensions UI API
  roadmap: Jan Core & Extensions SDK · status: duplicate · type: epic · janhq/jan
- jan #728
  epic: Jan Mobile
  roadmap: Jan Mobile · status: needs designs · type: epic · janhq/jan
- jan #1249
  epic: Jan Studio (users can finetune)
  roadmap: Jan Foundry · status: needs specs · type: epic · janhq/jan
- nitro #100
  feat: Add `pip` installation
  P1: important · type: feature request · janhq/nitro · Nitro Roadmap
- nitro #242
  Enhance E2E Testing: Add Additional Test Cases for Nitro on All Target OS (Windows, macOS, Ubuntu)
  roadmap: QA · type: enhancement

Planned 46

Committed to deliver

- jan #760
  epic: Keyboard Shortcuts, Search and Quick Switcher
  P2: nice to have · roadmap: Jan Core & Extensions SDK · type: epic · janhq/jan · Jan has Keyboard Shortcuts and Quick Swit...
- jan #1137
  docs: update Guide Using the Local Server
  P1: important · roadmap: Jan Docs · type: documentation · janhq/jan
- jan #1078
  epic: Jan Enterprise
  roadmap: Jan Enterprise · type: epic · janhq/jan
- jan #901
  feat: Give users Prompt suggestions
  roadmap: Assistants Framework · type: feature request · janhq/jan
- nitro #299
  feat: Nitro with Chatgptbox
  type: feature request · janhq/nitro

In Progress 23

This is actively being worked on

- jan #961
  Jan.AI mobile app
  P1: important · type: enhancement · janhq/jan · 2 reviewers
- jan #1299
  leave the app window active in the dock when closing on mac
  type: enhancement · janhq/jan · 1 reviewer
- jan #1337
  feat: Update model.json
  type: feature request · janhq/jan
- nitro #316
  epic: Support legacy hardware
  P0: critical · type: epic · janhq/nitro
- jan #1134
  docs: update Guide Customize Model Defaults
  roadmap: Jan Docs · type: documentation · janhq/jan
- jan #1232
  docs: openly disclose our analytics

In Review 3

This item is blocked on a PR/decision

- jan #1243
  bug: failed to fetch
  P0: critical · type: bug · janhq/jan · v0.4.5
- nitro #175
  Discussion: Nitro to automagically handl Model Operations?
  P1: important · type: discussion · janhq/nitro · Nitro Roadmap
- nitro #303
  feat: Nitro chat completion with image supporting local image path
  good first issue · type: feature request · janhq/nitro · Jan can See

- Working on data teams will require that you work in the same codebase as your peers
- This class aims to give you some experience with the following
  - Managing your work via tickets and a develop your team project via a Scrum
  - Working on branches to avoid conflicts
  - Code reviews and merging
- How github helps
  - Issues to break up your work into smaller pieces. Can comment, assign, etc.
  - Code reviews and Merge Requests
  - Discussions for team collaboration on the project as a whole
  - Projects on Github to keep your project organized

Helpful Tools (and some I think are really cool)

- Excalidraw (good for drafting/documenting architecture diagrams)
- Free academic license Lucidchart to help with ERDs, for one
- Motherduck (serverless duckdb - free 10gb)
- Neon (serverless postgres)
- Neo4j AuraDB (free hosted Neo4j database, with limits)
- MongoDB Atlas (free hosted MongoDB cluster)

Database - SQL review

DuckDB

MotherDuck